Harnessing the Potential of Machine Learning for Bioinformatics using Big Data Tools

M. Usman Ali
Department of Computer Science
COMSATS Institute of Information Technology
Sahiwal, Pakistan
usman.sani1439@ciitsahiwal.edu.pk

Shahzad Ahmad
Department of Computer Science
COMSATS Institute of Information Technology
Sahiwal, Pakistan
shahzadahmad@ciitsahiwal.edu.pk

Javed Ferzund
Department of Computer Science
COMSATS Institute of Information Technology
Sahiwal, Pakistan
jferzund@ciitsahiwal.edu.pk

***Abstract*—Advancement in the sequencing technology has resulted in the generation of large amount of bioinformatics data that need to be analyzed in short time. Traditional techniques cannot cope with the speed and size of data generation. New platforms, tools and techniques need to be explored for data analysis that can meet the time and space challenge. Big data tools and techniques address the issues of size, scalability and performance. In this paper, we present the potential of machine learning techniques implemented on Hadoop and Spark for analyzing bioinformatics data. First, we summarize the recent work in this area and then discuss the future research directions and opportunities.**

Keywords-Big Data, Machine Learning, Hadoop, MapReduce, Spark, Bioinformatics, Microarray, Gene, DNA, RNA, Proteomics

I.INTRODUCTION

Bioinformatics domain includes DNA, RNA, protein and genomics data in the heterogeneous network such as gene-gene interaction, protein-protein interaction and gene-disease interaction. During the last few years, this data has enormously increased in volume and need to be processed in a well-organized manner to reduce the execution time and space requirements. Usually microarray data analysis for gene selection and classification require closely related genes in the proper way and also DNA sequencing data is very important for analysis. Many tools have designed for bioinformatics data analysis such as blast, EMBOSS, bioperl, babel and Modeller. All these tools work on small datasets and do not illustrate any performance on large datasets. However, there is no designed appropriate benchmark that can fit to all of these problems. So, there is need to design finest benchmark to manage, evaluate, store and analyze large datasets produced in many experiments of sequencing, proteomics and genomics.

A large amount of data has been produced in engineering, biological and computational fields. More recently, big data has been introduced to analyze, manage and store the large datasets such as google and yahoo data. In situations where data cannot be controlled by ordinary techniques big data can helps.

Big data handles basically five V's i.e. volume, variety, velocity, veracity and potential value. Big data techniques contain data mining (association rule learning), crowdsourcing, big cloud for computers, linear and multiple regression analysis and investigation of social networks. For parallel processing of large datasets, Hadoop platform has been designed that includes modules HDFS (Hadoop Distributed File System), MapReduce, Pig, Hive, HBase, yarn and Spark. Hadoop is an open source JAVA written platform that gives distributed and parallel processing and storage capability for large datasets through many clusters. HDFS provides distributed storage with the help of nodes of clusters using one Name node (master) and one or multiple Data nodes (slaves) for all other Hadoop modules [1]. MapReduce is a framework for parallel processing of large data sets in reliable and fault-tolerant way using one Job Tracker (master) and one Task Tracker (slave) in the key and value pair with both Map and Reduce tasks [1]. Hive is data warehouse framework that provides HQL (Hive Query Language) like SQL (Structured Query Language) interface for ad-hoc queries and summarization in batch processing environment [1]. Pig is scripting language also for batch processing and HBase provides random read write access and real time processing of big data [1]. Hadoop is the superlative platform for large biological data processing of DNA, RNA, protein and genomics and also plays an important role in microarray data analysis and read mapping.

This paper was submitted for review on 23 October 2016.
M. Usman Ali is with the Department of Computer Science, COMSATS Institute of Information Technology, Sahiwal, 57000 Pakistan (e-mail: usman.sani1439@ciitsahiwal.edu.pk)
Shahzad Ahmad is with the Department of Computer Science, COMSATS Institute of Information Technology, Sahiwal, 57000 Pakistan (e-mail: shahzadahmad@ciitsahiwal.edu.pk)
Javed Ferzund is currently working in the Department of Computer Science, COMSATS Institute of Information Technology, Sahiwal, 57000 Pakistan (e-mail: jferzund@ciitsahiwal.edu.pk)

Many machine learning techniques and algorithms are used for classification and analysis of datasets. Machine learning classification and regression algorithms include decision tree, naïve bays, logistic regression, SVM (Support Vector Machine), gradient boosted tree, random forest, generalized linear model, linear regression. Clustering algorithms are K-means, Fuzzy K-means, power iteration, spectral clustering and CluStream. Machine learning also consist of association rule mining and deep learning. All of these machine learning algorithms and techniques are used in traditional computational processes. Machine learning also, has great importance in the field of big data. Almost these techniques are used in Hadoop big data framework such as in the MapReduce and Spark. MapReduce uses Mahout Library (written in Java) and Spark uses Mlib library (written in Java, Python, and Scala) for implementation of machine learning techniques. Some of machine learning techniques used in Hadoop big data are mentioned in the Table 1.

The objectives of this study are:

- To explore the Machine Learning techniques used in Bioinformatics
- To analyze the capabilities o f Machine Learning techniques implemented on Big Data Platform
- To present the future research opportunities for using Machine Learning techniques along with Big Data platform to analyze the bioinformatics data
- To explore the Performance comparison of existing Algorithms

The rest of the paper is structured as follows: Section II describe the related work. Section III represent the work in Bioinformatics along with Machine Learning. Section IV represent the use of ML tools, algorithms and techniques in the field of Bioinformatics using Big Data Hadoop. Section V describe the discussion related to our work. Section VI concludes the paper with opportunities for further research work in Bioinformatics along with Big Data Hadoop.

II.RELATED WORK

A. Machine Learning in Bioinformatics

In the past, machine learning techniques have been used in bioinformatics domain for microarray data analysis. Hernandez et al. [2] proposed computational approach for selection and classification of genes, in which they classified the genes with SVM (Support Vector Machine) classifier with the help of genetic algorithm by using leukemia, colon cancer and lymphoma datasets from NCBI resulting higher accuracy. Leu et al. [3] have developed analysis of microarray data with sampling, in which genes are classified into three groups based on expression level. After removing the unnecessary groups, subsets are made by using sampling. Then irrelevant subsets are removed with the help of classification accuracy determined by kNN algorithm and χ2-test is used to find relevant genes resulting in best classification accuracy with fewer genes by using three bioinformatics datasets from NCBI. Lee et al. [4] have proposed novel gene selection approach for microarray, in which GADP (Genetic Algorithm with Dynamic Parameter setting) is used with the χ2-test for gene selection and also SVM (support vector machine) is used for efficiency verification of genes Resulting in best classification accuracy with fewer genes by using six bioinformatics datasets from NCBI. Kumar et al. [5] have proposed Fuzzy kNN algorithm classification, providing 100% accuracy to select the genes with t-test and classify the genes using kNN by using leukemia and breast cancer datasets.

Table 1: Overview of Machine Learning techniques used in Hadoop Big Data

Machine Learning (Techniques and Algorithms)	**MapReduce (Mahout library)**	**Spark (Mlib library)**
NB (Naïve Bayes Bayesian Algorithm)	Yes	Yes
GBT (Gradient Boosted Tree Ensemble Algorithm)	Yes	Yes
Streaming K-means Clustering	Yes	Yes
SVM (Support Vector Machine)	Yes	Yes
K-means Clustering	Yes	Yes
Adaptive Model Rules	No	No
GLM (Generalized Linear Model)	No	Yes
kNN (Instance based Algorithm)	Yes	Yes
LR (Logistic Regression)	Yes	Yes
Deep Learning	Yes	Yes
Random Forest (Ensemble Algorithm)	Yes	Yes
k-Median Clustering	Yes	Yes
Association Rule Mining (Apriori Algorithm)	Yes	Yes
Decision Tree	Yes	Yes
Linear Regression	Yes	Yes

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

B. *Machine Learning in Big Data Hadoop Platform*

Machine learning techniques are being used in big data Hadoop platform since last in the few years. Ye et al. [6] have developed Stochastic GBDT (Gradient Boosted Decision Trees) learning algorithm for machine learning. It presents two methods for improvement of training time individual trees that produces exact stochastic GBDT models that are implemented on MapReduce and then implemented on Hadoop with MPI (Message Passing Interface). Dai et al. [7] have developed MapReduce based application of Decision Tree C 4.5 that reduces the time and memory requirements and results in efficient and scalable method for large datasets. Venkataraman et al. [8] have developed Generalized Linear Model in SparkR with the help of R tool for large datasets. R is statistical tool that provides R package in Hadoop framework such as SparkR for Spark and RHIPE for MapReduce. RHIPE is used to calculate the correlation matrix on gene expression data [9]. OANCEA et al. [10] have performed LR (Linear Regression) using statistical tool R and Hadoop framework by using open source Rhadoop library for large datasets and least squares solution for the linear regression problem have been conveyed in terms of map-reduce framework.

C. *Machine Learning in Bioinformatics using Big Data Hadoop Platform*

Traditionally, there are many methods of gene selection for microarray data analysis such as GADP, SVM and supervised clustering. All of these methods are better to some extent but not good for scalability. Recently, machine learning techniques are used in Bioinformatics domain using Big Data framework. A.K.M. Tauhidul Islam et al [11] have developed Microarray data analysis with the help of Hadoop MapReduce platform, in which map task find out BW (Between-groups to Within-groups sum of square) ratio value for every gene. BW measures find out degree of variance between gene expression values. After finding potential gene subset, it uses kNN classifier algorithm on gene list for accuracy. By running multiple parallel kNN algorithms known as MRkNN for finding top-k genes by using 4 real and 3 synthetic bioinformatics datasets, better results are obtained in terms of accuracy and scalability. Kumar et al. [12] have also developed a method in which statistical test ANOVA (Analysis of Variance) is used for gene selection and kNN algorithm is used to classify the features resulting in better speedup and scalability by using NCBI datasets. Ray et al. [13] have proposed Spark framework for microarray data analysis in such a way that feature selected with sf-ANOVA and genes are classified with machine learning techniques such as Logistic Regression and Naïve Bays resulting in best accuracy, scalability and speedup as compared to all traditional methods.

III. BIOINFORMATICS

Bioinformatics consists of multiple heterogeneous networks of DNA, RNA, protein and genome and their interaction in multiple ways. Machine learning techniques perform a significant role in the field of Bioinformatics in areas like gene prediction, microarray data analysis, sequence alignment, pattern identification, protein-protein interaction prediction and SNP (Single Nucleotide Precision). Multiple Machine Learning techniques and algorithms are used in gene selection and classification for microarray data analysis. For large datasets, Some of Machine Learning techniques are used in Bioinformatics using Big Data Hadoop framework. However, many techniques are still not being used in Bioinformatics with Big Data Hadoop. By using these techniques, we can expect better performance, scalability, efficiency, accuracy and speedup. Basically, this is the main focus.

After gene selection and classification for microarray data analysis using Machine Learning techniques with Hadoop platform, we can efficiently perform Katz (link formulation method), CATAPULT (positive unlabeled method) and IMC (Inductive Matrix Calculation) for gene-disease association to find how much a gene is associated with specific disease [14]. Also, can well perform sequence alignment and GATK pipeline after microarray data analysis using ML (Machine Learning) with Hadoop [15]. However, ML with Hadoop also show an imperative role in Bioinformatics and engineering fields in terms of different scenario.

Machine Learning techniques used in Hadoop context are presented in Table 1. These almost can be used in Bioinformatics domain with Big Data Hadoop for designing best resultant benchmark for large datasets. By using these techniques, we can address the problems of time, space, scalability and speedup.

IV.MACHINE LEARNING TOOLS AND TECHNIQUES

There are many tools of Machine Learning such as Hadoop Mahout Library (uses MapReduce), Caffe, Mlib Library (uses Apache Spark), WEKA, Neon, Torch and ConvNetJS. Most of Machine Learning tasks are implemented via Hadoop libraries such as Mahout and Mlib. Mahout offers performance and scalable features for analysis of large datasets using Machine Learning in the perspective of clustering and classification. Mlib is a Spark Machine Learning library which includes many algorithms such as Naïve Bays and Decision Tree etc. it gives best scalability, speedup and performance parameters as compared to MapReduce. Caffe provides Deep Learning platform for Machine Learning tasks in terms of Speed and used in making different models for analysis in the perspective of learning tasks. WEKA is Data Mining and Machine learning tool. Torch is a platform for GPU and provides flexibility and speedup.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 2: Overview of Machine Learning techniques used in Bioinformatics using Big Data Hadoop

Machine Learning (Techniques and Algorithms)	**MapReduce (Mahout library)**	**Spark (Mlib library)**	**Bioinformatics using Big Data (Bioinformatics+Hadoop)**
NB (Naïve Bayes Bayesian Algorithm)	Yes	Yes	Yes
GBT (Gradient Boosted Tree Ensemble Algorithm)	Yes	Yes	No
Streaming K-means Clustering	Yes	Yes	No
SVM (Support Vector Machine)	Yes	Yes	Yes
K-means Clustering	Yes	Yes	No
Adaptive Model Rules	No	No	No
GLM (Generalized Linear Model)	No	Yes	No
kNN (Instance based Algorithm)	Yes	Yes	Yes
LR (Logistic Regression)	Yes	Yes	Yes
Deep Learning	Yes	Yes	No
Random Forest (Ensemble Algorithm)	Yes	Yes	Yes
k-Median Clustering	Yes	Yes	No
Association Rule Mining (Apriori Algorithm)	Yes	Yes	No
Decision Tree	Yes	Yes	No
Linear Regression	Yes	Yes	Yes

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A lot of Machine Learning techniques are used today mentioned in Table 1. Some are used in Bioinformatics; some are in Big Data Hadoop and some are in Bioinformatics with Big Data Hadoop. Naïve Bayes is used in training and testing of data, making models in statistics and having major focus on classification of datasets. Gradient Boosting solves the problems of regression and classify the datasets with Decision Tree model. Random Forest behaves as GBT (Gradient Boosted Tree) but performance is best with GBT. SVM (Support Vector Machine) combined with GA (Genetic Algorithm) provides more accuracy for large datasets. Fuzzy kNN algorithm performs better than kNN to classify large datasets. Association Rule Mining give the relationship between different variables.

There are many Machine Learning Techniques and Algorithms that are implemented in Bioinformatics using Big Data Hadoop framework. Some of these are Naïve Bayes (Bayesian Algorithm), SVM (Support Vector Machine), kNN (Instance based Algorithm), Logistic Regression, Random Forest (Ensemble Algorithm) and Linear Regression that are mentioned in Table 2. Naïve Bayes and Logistic Regression used for microarray classification in Apache Spark for large datasets acquire best scalability. kNNs are implemented in Hadoop MapReduce platform for microarray feature classification for big datasets for better performance. SVM (Support Vector Machine) also provides best results for gene selection/prediction for microarray data analysis in Hadoop. Better result appeared with Random Forest in Bioinformatics using Big Data Hadoop.

Some Machine Learning Techniques are not implemented in Bioinformatics using Big Data Hadoop such as Gradient Boosted Tree (Ensemble Algorithm), Streaming K-means clustering, K-means clustering, Adaptive Model Rules, GLM (Generalized Linear Model), Deep Learning, K-median clustering, Association Rule Mining (Apriori Algorithm) and Decision Tree that are mentioned in Table 2. By using these ML techniques and algorithms in Bioinformatics using Big Data Hadoop, we can achieve superlative Performance, Accuracy, Scalability, Reliability, Speedup and Efficiency by reducing the need for time and storage.

A. Performance Comparison of Existing Algorithms

By implementing NB (Naïve Bayes) Technique in Spark framework for Bioinformatics (Microarray) data, better Accuracy and Performance has been obtained as compared to Hadoop MapReduce framework and conventional Techniques [17]. Naïve Bayes provides better Accuracy and Scalability than Logistic Regression in Spark framework. Apache Spark offers best Performance by reducing training time when kNN Algorithm is implemented on Spark. By implementing SVM (Support Vector Machine) Algorithm in Hadoop MapReduce framework for Bioinformatics (Microarray) data, better Accuracy has been achieved than kNN. SVM implemented on MapReduce gives better Performance by decreasing training time than using SVM on traditional tools [17].

Scalability of SVM is low as compared to kNN Algorithm when it is implemented in MapReduce framework. By implementing kNN (k Nearest Neighbor) Algorithm in Hadoop MapReduce framework for Bioinformatics (Microarray) data, better Scalability has been achieved as compared to SVM and GADP (Genetic Algorithm with Dynamic Programming). Similarly, Accuracy of kNN is less than SVM. Implementation of kNN in MapReduce provides best Performance by decreasing communication cost than implementation of kNN in traditional tools.

By implementing LR (Logistic Regression) Technique in Spark framework for Bioinformatics (Microarray) data, better Accuracy can be obtained than Hadoop MapReduce framework. Implementation of LR in Spark provides lower Accuracy and Scalability as compared to NB. By implementing Random Forest Algorithm in Hadoop MapReduce framework for Bioinformatics (Microarray) data, Accuracy can be maintained in the presence of missing data. Scalability of Random Forest is better than other algorithms. It provides best Performance by using many-many model. Implementation of Linear Regression in Hadoop MapReduce framework for Bioinformatics (Microarray) data also provides better Accuracy, Scalability and Performance. Linear Regression is most widely used ML Technique that runs fast. Table 3 explains Performance Comparison of existing algorithms.

V.DISCUSSION

In the recent past, With the passage of time, use of Machine Learning tools and techniques along with are used bioinformatics using Big Data Hadoop have gained popularity in the bioinformatics domain. SVM (Support Vector Machine) and kNN are implemented in Bioinformatics using Big Data Hadoop for genome datasets. We can use SVM for protein datasets such as to find protein-protein interaction networks. Logistic Regression and Naïve Bayes techniques are also used in Hadoop for gene classification for microarray data. These techniques give better performance when we use them for proteomics. Linear Regression also can be used for protein datasets. By using the state of the art machine learning techniques implemented on Hadoop and Spark for DNA, RNA and protein datasets, we can expect best Performance, Accuracy, Speedup and Scalability. A lot of research potential exists in this area.

The ML techniques and algorithms that are used in Bioinformatics using Hadoop can also be used in DNA and RNA Sequence Alignment to achieve best performance for some analysis. These techniques and algorithms can also be used in GATP pipeline for best read mapping Scalability and Performance. With the help of using these tools, techniques and algorithms in the domain of Bioinformatics along with Big Data Hadoop demonstrate extreme potential in our work.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 3: Performance Comparison of Existing Algorithms

Machine Learning (Techniques and Algorithms)	**Dataset**	**Implementation Platform (Hadoop)**	**Accuracy**	**Scalability**	**Performance and Comparison with Traditional Algorithms**
NB (Naïve Bayes Bayesian Algorithm)	Bioinformatics	Apache Spark	Using Spark, Better than MapReduce and conventional. Better than LR	Same as like MapReduce. Better than LR	Using Spark, 100 times faster than MapReduce. Training time decreases and best Performance than traditional NB
SVM (Support Vector Machine)	Bioinformatics	MapReduce	Better than kNN	Less than kNN	Training time decreases, best Performance and reduce computational complexity than traditional SVM
kNN (Instance based Algorithm)	Bioinformatics	MapReduce	Less than SVM	Better than SVM, BPNN and GADP Algorithms	Better Performance and decreases communication cost than traditional kNN
LR (Logistic Regression)	Bioinformatics	Apache Spark	Using Spark, Better than MapReduce and conventional. Less than NB	Same as like MapReduce. Less than NB	Using Spark, 100 times faster than MapReduce. Execution time decreases and best Performance than traditional LR
Random Forest (Ensemble Algorithm)	Bioinformatics	MapReduce	Maintains Accuracy when there is missing data	Scalability better than others Algorithms	Best using many-many model
Linear Regression	Bioinformatics	MapReduce	Best than traditional Techniques	Work fine on high-dimensional, sparse dataset	Most widely used ML Technique that runs fast

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

ML tools used with Big Data Hadoop have produced surprising results in Bioinformatics. For gene classification, fuzzy kNN algorithm in MapReduce platform gives better results for scalability and accuracy for large datasets than by using kNN. We can implement Deep Learning and Decision Tree techniques in Apache Spark for microarray data analysis to get best Accuracy.

We can reduce the time and space constraints to solve Bioinformatics problems by implementing all of these techniques on Hadoop and Spark.

Scalability of NB (Naïve Bayes) Technique can be improve by implementing in Apache Spark framework for large Bioinformatics dataset. There is a great need to improve Scalability of SVM (Support Vector Machine) Algorithm that is implemented in MapReduce. We can improve the Accuracy, Scalability and Performance of SVM Algorithm by implementing in Spark for large Bioinformatics dataset. There is a great need to improve Accuracy of kNN (k Nearest Neighbor) Algorithm that is implemented in MapReduce. We can improve the Accuracy, Scalability and Performance of kNN Algorithm by implementing in Spark for large Bioinformatics dataset. Accuracy and Scalability of LR (Logistic Regression) Technique can be improved by implementing in Apache Spark framework because LR Accuracy and Scalability have less than NB in Spark. Accuracy, Scalability and Performance of Linear Regression and Random Forest Algorithm can be improved by implementing that Algorithms in Apache Spark framework. The Performance comparison is given in *Table 3*.

VI.CONCLUSION

In this paper, we describe the many Machine Learning Tools, Techniques and Algorithms for Bioinformatics using Big Data Hadoop for large datasets expected to preeminent results by reducing the need for time and space limits. We distinguish the implementation of ML techniques and algorithms that are used in Bioinformatics domain using Big Data Hadoop mentioned in Table 2. With the usage of these ML techniques such as Deep Learning, Streaming K-means clustering, Decision Tree, Adaptive Model Rules, GLM (Generalized Linear Model), Gradient Boosted Tree (Ensemble Algorithm), K-median clustering, Association Rule Mining (Apriori Algorithm) and K-means clustering, we conclude that these gives best Accuracy, Performance, Scalability and Speedup directed to big ML potential for Bioinformatics using Big Data.

Adaptive model rules are not used in Hadoop platform. Generalized linear model are not used in MapReduce. We can implement Adaptive model rules in Hadoop MapReduce and Apache Spark to attain superlative results. Also, we can use GLM in MapReduce platform for best Performance analysis.

ACKNOWLEDGMENT

The authors would like to express thanks to Abbas Rehman and Atif Sarwar Department of Computer Science, COMSATS Institute of Information Technology Sahiwal, Pakistan for their visionary suggestions and beneficial contribution to support this work.

REFERENCES

[1] Taylor, Ronald C, "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics," in *Bioinformatics Open Source Conference (BOSC)*, Boston, MA, USA, 2010.

[2] Jose Crispin Hernandez Hernandez , Jin-Kao Hao, B´ eatrice Duval;, "A Genetic Embedded Approach for Gene Selection and Classification of Microarray Data," in *EvoBio*, Verlag Berlin Heidelberg, 2007.

[3] Yungho Leu, Chien-Pang Lee, and Hui-Yi Tsai, "A Gene Selection Method for Microarray Data Based on Sampling," in *ICCCI*, Verlag Berlin Heidelberg, 2010.

[4] Chien-Pang Lee , Yungho Leu, "A novel hybrid feature selection method for microarray data analysis," *Applied Soft Computing*, vol. 11, no. 1, pp. 208-213, January 2011.

[5] Mukesh Kumar, Santanu Ku. Rath, "Microarray Data Classification using Fuzzy K-Nearest Neighbor," 2014.

[6] Jerry Ye, Jyh-Herng Chow, Jiang Chen, Zhaohui Zheng, "Stochastic Gradient Boosted Distributed Decision Trees," in *CIKM*, Hong Kong, China, 2009.

[7] Wei Dai, Wei Ji, "A MapReduce Implementation of C4.5 Decision Tree Algorithm," *International Journal of Database Theory and Application*, vol. 7, pp. 49-60, 2014.

[8] Venkataraman et al., "SparkR: Scaling R Programs with Spark," in *SIGMOD*, San Francisco, CA, USA, 2016.

[9] Wang et al., "Optimising parallel R correlation matrix calculations on gene expression data using MapReduce," in *BMC Bioinformatics*, 2014.

[10] OANCEA, Bogdan, "LINEAR REGRESSION WITH R AND HADOOP," *Challenges of the Knowledge Society*, pp. 1007-1012, 2016.

[11] A.K.M. Tauhidul Islam, Byeong-Soo Jeong, A.T.M. Golam Bari, Chae-Gyun Lim, Seok-Hee Jeon, "MapReduce based parallel gene selection method," *Applied Intelligence*, vol. 42, no. 2, pp. 147-156, 2015.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[12] Mukesh Kumar, Nitish Kumar Rath, Amitav Swain, Santanu Kumar Rath, "Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor," in *IMCIP*, 2015.

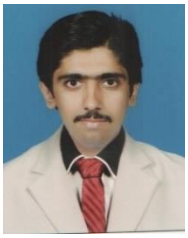
[13] Ransingh Biswajit Ray, Mukesh Kumar, Santanu Kumar Rath, "Fast Computing of Microarray Data Using Resilient Distributed Dataset of Apache Spark," in *Recent Advances in Information and Communication Technology*, vol. 463, Springer International Publishing, 2016, pp. 171-182.

[14] Nagarajan Natarajan, Inderjit S. Dhillon, "Inductive Matrix Completion for Predicting Gene-Disease Association," *Oxford*, vol. 30, no. 12, 2014.

[15] Hamid Mushtaq, Zaid Al-Ars, "Cluster-Based Apache Spark Implementation of the GATK DNA Analysis Pipeline," in *BIBM*, Washington, USA, 2015.

[16] Sara Landset, Taghi M. Khoshgoftaar, Aaron N. Richter, Tawfiq Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," *Journal of Big Data*, 2015.

[17] S. R. Pakize and A. Gandomi, "Comparative Study of Classification Algorithms Based on MapReduce Model," International Journal of Innovative Research in Advanced Engineering (IJIRAE), vol. 1, no. 7, pp. 251-254, August 2014.



M. Usman Ali is a Lab Engineer at Department of Computer Science, COMSATS Institute of Information Technology Sahiwal, Pakistan. He received BS (IT) degree from Govt. College University Faisalabad, Pakistan in 2013. Currently, he is a scholar of MS (CS) session 2015-2017 in COMSATS Institute of Information Technology Sahiwal, Pakistan. His main research interests include Big Data Analytics and Machine Learning. Particularly, he is interested in applications of Big Data in the Bioinformatics field. Currently, he is working with the Big Data Analytics Research Group at COMSATS Institute Sahiwal.

Shahzad Ahmad is a Research Associate at Department of Computer Science, COMSATS Institute of Information Technology Sahiwal, Pakistan. He received BS (CS) degree from COMSATS Institute of Information Technology Sahiwal, Pakistan in 2015. Currently, he is a scholar of MS (CS) session 2015-2017 in COMSATS Institute of Information Technology Sahiwal, Pakistan. His main research interests include Big Data Analytics and Machine Learning. Particularly, he is interested in applications of Big Data in the Bioinformatics field. Currently, he is working with the Big Data Analytics Research Group at COMSATS Institute Sahiwal.

Dr. Javed Ferzund is an associate professor at Department of Computer Science, COMSATS Institute of Information Technology, Sahiwal, where he served as Head of Department from 2013-2015. He received PhD degree from Graz University of Technology, Austria in 2009. His main research interests include Big Data Analytics, Internet of Things and Machine Learning. Particularly, he is interested in applications of IoT and Big Data in the Agro-Informatics and Bioinformatics fields. Currently, he is leading the Big Data Analytics Research Group at COMSATS Institute Sahiwal.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500